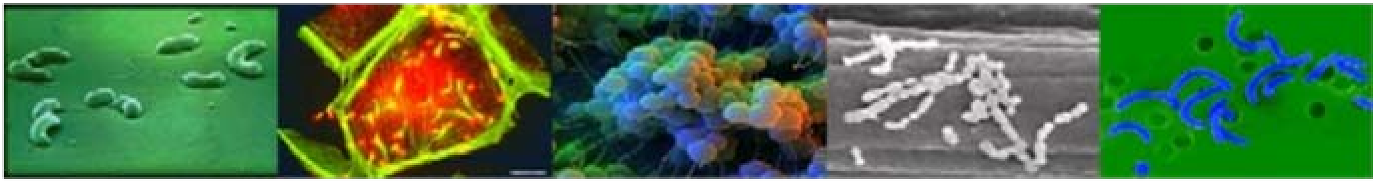
NMPDR National Microbial Pathogen Data Resource Center



Rapid Propagation of Annotations

by

Ross Overbeek

THE UNIVERSITY OF CHICAGO

FELLOWSHIP FOR INTERPRETATION OF GENOMES

National Institute of Allergy and Infectious Diseases
National Institutes of Health

NCSA

ARGONNE NATIONAL LABORATORY

Construct a Set of Ortholog Families (4 tables)

- Family[family_id,genome,gene_id]
- FamilyFunction[family_id,function]
- DnaSeq[genome,gene_id,DNA-sequence]
- ProteinSeq[genome,gene_id,protein-sequence]

Add Support for Large Repetitive Elements (LRE) (Call the result a PSF)

These include prophages and transposition events

- LargeRepetiveElement[LREid,genome,gene_id]
- LRE_function[LREid,function]

Basic Specs:

```
propagate_annotations PSF contigs  GFF3-annotations UpdatedPSF [parameters]
```

where input would be

PSF would a directory containing six files, each containing a tab-separated table

contigs is a fasta file of contigs

and output would be

GFF3-annotations for the new genome

UpdatedPSF an updated version of the PSF

Steps in Processing

1. Propagate Families (producing a set of CDSs and RNAs)
2. Propagate the LREs
3. Use CDSs from Steps 1 as a training set for GLIMMER, and call genes
4. Remove genes from GLIMMER calls when they match those from step 1 or overlap those from step 2
5. Blast remaining GLIMMER hits against a non-redundant protein DB
6. Remove overlaps (use similarities to resolve priorities)
7. Generate GFF3 output

Add Support for Subsystem Propagation (5 tables)

- RoleInSubsystem[Subsystem,FunctionalRole]
- FamilyPlaysRole[family_id,FunctionalRole]
- LREplaysRole[LREid,FunctionalRole]
- ActiveVariant[Subsystem,genome,variant-id]
- VirulenceRelated[Subsystem]

Summary

1. Construct Accurate Families and LREs (call the result a PSF)
2. Build a Tool to Propagate and extend PSFs
3. Add Support for Propagation of Subsystems
4. Take a new genome and (within 2 days)
    a) Call the genes
    b) Assign reasonably accurate functions
    c) Produce a metabolic reconstruction
    d) Construct an inventory of virulence factors